DOCUMENT RESUME

ED 193 300 TM 800 615

AUTHOR Haebara, Tomokazu
TITLE Equating Logistic Ability Scales by a Weighted Least Squares Method. Iowa Testing Programs Occasional Papers, Number 27.
REPORT NO ITPOP27
PUB DATE Aug 80
NOTE 21p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Discriminant Analysis; *Equated Scores; Goodness of Fit; Grade 7; Grade 9; *Intelligence Tests; Item Analysis; Junior High Schools; *Latent Trait Theory; Least Squares Statistics; *Mathematical Models; *Measures (Individuals)
IDENTIFIERS *Optimal Scaling

ABSTRACT

This study develops a weighted least squares method for transforming a logistic scale in such a way that the estimates of ability parameters on the tranformed scale are as comparable as possible with those on another scale. This scale tranformation process is referred to as equating of scales. Equating is an important procedure in studies involving the logistic model and the latent trait model in general; the equating method presented deals with situations in which there are two separately scaled tests that have only some items in common (vertical equating or horizontal equating) or all items in common (horizontal equating). The necessity for equating in the later case arises when two scales have been derived by using two different groups of examinees which may have different characteristics pertinent to the results of scaling. In the former case (only some items common), the groups of examinees may or may not be overlapped. As one of the optimum estimation procedures, a weighted least squares method is introduced and applied to an achievement scale and an intelligence test. An advantage of the method over previously proposed methods is that highly discriminating items tend to affect equating more strongly than less discriminating items. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193300

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IOWA TESTING PROGRAMS OCCASIONAL PAPERS

Number 27 - August 1980

Equating Logistic Ability Scales by a Weighted Least Squares Method

Tomokazu Haebara

The University of Iowa

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. Haebara
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The material presented here is automatically superseded upon formal publication.

TM 800615

ABSTRACT

A general procedure for equating logistic ability scales that involves the use of a loss function and an optimization process is formulated. As one of the optimum estimation procedures, a weighted least squares method is introduced and applied to an achievement scale and an intelligence scale. An advantage of the method over previously proposed methods is that highly discriminating items tend to affect equating more strongly than less discriminating items.

ACKNOWLEDGMENT

I would like to thank Drs. William E. Coffman and D. R. Divgi of the University of Iowa and Dr. Sukeyori Shiba of the University of Tokyo for their valuable comments on earlier versions of the paper. I am also indebted to Shin-ichi Mayekawa (numerical analysis), Jim Keasler (data processing), and Kary Jacobs (typing).

INTRODUCTION

The purpose of this study is to develop a weighted least squares method for transforming a logistic scale in such a way that the estimates of ability parameters on the transformed scale are as comparable as possible with those on another scale. In the present paper, such a scale transformation process is referred to as equating of scales.

Equating is an extremely important procedure in studies involving the logistic model and the latent trait model in general. To mention a few examples, Shiba's stratified adaptive test of verbal ability is based on successive vertical equatings of thirteen logistic scales that jointly cover the ability levels of preschool children through adults (Shiba, 1978; Shiba, Noguchi, & Haebara, 1978; Shiba, Noguchi, & Ohama, 1980). Haebara's (1979) study on item bias largely depends on a horizontal equating of logistic scales for two racial groups.

The equating method to be presented in this paper deals with situations in which there are two separately scaled tests that have only some items in common (vertical equating or horizontal equating) or all items in common (horizontal equating). The necessity for equating in the latter case (all items common) arises when two scales have been derived by using two different groups of examinees which may have different characteristics pertinent to the results of scaling. In the former case (only some items common), the groups of examinees may or may not be overlapped.

A conceptually simple way of equating in the situations described above is to pool all items in the two tests and all examinees in the two groups, and to reestimate examinee ability and item characteristic curve parameters simultaneously, using the enlarged set of data. However, this procedure is often prohibitively expensive, not only because it involves a large number of examinees and/or items but also because it wastes results available from the previous scaling. Hence, it is desirable to develop an alternative method for equating that utilizes the previously obtained scales.

FORMULATION

Logistic Model

In the logistic test model, the probability of an examinee with ability parameter θ giving a correct answer to item g is mathematically stated as

$$P_g(\theta) = c_g + \frac{1 - c_g}{1 + \exp\left[-Da_g(\theta - b_g)\right]}, \qquad (1)$$

where a_g, b_g, and c_g are parameters describing item g, and D is a constant assuming the value of 1.7 (see Birnbaum, 1968; Hambleton, Swaminathan, Cook, Eigner, & Gifford, 1978). Equation 1 can be conceived of as representing the following three different logistic models: (a) Three-parameter model, where each of the three parameters is to be determined for each item; (b) Two-parameter model, where c_g is zero for all items; and (c) One-parameter model, where a_g is a constant across all

items and c_g is zero for all items. Although the ensuing discussion will focus on the three-parameter model, the method to be presented later is applicable to the other logistic models as well.

Principle of Equating

Let two logistic scales to be equated be Scale 1 and Scale 2. By the nature of the logistic model, the two scales have arbitrary units and origins. Therefore, for an ability parameter θ_1 on Scale 1 to be comparable with ability parameters on Scale 2, it must be transformed to an equated value θ_{12} by

$$\theta_{12} = k\theta_1 + d, \tag{2}$$

where k and d are the equating coefficients. (Any nonlinear transformation of the scale violates the basic assumption of the logistic model by causing non-logistic item characteristic curves.)

Then a perfect equating implies that

$$P_{g,1}(\theta_1) = P_{g,2}(\theta_{12}), \qquad -\infty < \theta_1 < \infty, \tag{3}$$

for all common items g (g = 1, 2,..., m, say). Here the second subscript to P_g refers to the scale it is defined on. Thus the equating problem in the present context is to find the values of k and d in Equation 2 that satisfy Equation 3 for all of the m common items.

In practice, however, Equation 3 never holds for all common items because of sampling errors and a possible lack of fit of the model. Therefore, the actual task is to estimate k and d so that Equation 3 holds as nearly as possible. An objective and valid way of solving the

problem is to define a criterion function that reflects the degree of departure of an equating result from the relationship stated in Equation 3 and to find such values of k and d as to minimize the criterion function.

For item g (g = 1, 2,..., m) and examinee a (a = 1, 2,..., N), let $e_{ga,1}$ be defined by

$$e_{ga,1} = P_{g,1}(\theta_{1,a}) - P_{g,2}(\theta_{12,a}). \tag{4}$$

Then, by choosing an appropriate loss function L that evaluates the loss caused by the magnitude of $e_{ga,1}$, the equating error QI_g for item g can be written

$$QI_g = \sum_{a=1}^{N} L(e_{ga,1}).$$

Furthermore, the total equating error Q is defined by

$$Q = \sum_{g=1}^{m} QI_g. \tag{5}$$

If the values of k and d minimizing Q are found, they are the optimum equating coefficients for the loss function L.

Previously Developed Methods

Although several equating methods have been developed and used in the past few years, none of them have adopted the optimization procedure described above. Those methods are characterized by the use of the moments of the distributions of item parameters a_g and b_g on the two scales. The basic idea underlying those methods is that, when the equating is defined by Equation 2, the a_g and b_g parameters on Scale 1 must be transformed by

$$a_{g,12} = a_{g,1}/k$$

and

$$b_{g,12} = kb_{g,1} + d$$

to be comparable with the item parameters on Scale 2. Here the second subscript to the item parameter refers to the scale it is defined on.

For example, Marco (1977) and Haebara (1979) estimated k and d by

$$k = (\text{SD of } b_{g,2})/(\text{SD of } b_{g,1}) \tag{6}$$

and

$$d = (\text{Mean of } b_{g,2}) - k(\text{Mean of } b_{g,1}), \tag{7}$$

where the means and SD's are computed by using all common items. Shiba (1978) proposed two different ways of estimating k:

$$k = (\sum_{g=1}^{m} a_{g,1})/(\sum_{g=1}^{m} a_{g,2});$$

and

$$k = (\sqrt{1 + \alpha^2} - 1)/\alpha$$

if

$$\alpha = (2\sum_{g=1}^{m} a_{g,1}a_{g,2})/\sum_{g=1}^{m} (a_{g,2}^2 - a_{g,1}^2) > 0,$$

and

$$k = \alpha/(1 - \sqrt{1 + \alpha^2})$$

if $\alpha < 0$. More recently, Douglass (1980) claimed that the estimate of k given by the average of $a_{g,1}/a_{g,2}$ provided a reasonable equating.

As Douglass (1980) admitted, however, the best method of combining the information contained in the a_g and b_g for estimating k and d is not

obvious as far as such heuristic approaches are concerned. If the optimization procedure described in the preceding section is employed, this kind of ambiguity disappears once an appropriate loss function has been selected by the researcher. In the next section, one of the optimum estimation procedures which employs the weighted least squares criterion is discussed.

METHOD

Criterion Function

Let the loss function L be defined by

$$L(e_{ga,1}) = e_{ga,1}^2,$$

where $e_{ga,1}$ is given by Equation 4. Then Q in Equation 5 becomes

$$Q = \sum_{g=1}^{m} \sum_{a=1}^{N} e_{ga,1}^2.$$

If we construct a relative frequency distribution h_1 of θ_1 by dividing the range of θ_1 into some n_1 small intervals with the middle points $\theta_{1,i}$ ($i = 1, 2, \ldots, n_1$), minimizing Q is approximately equal to minimizing Q_1 defined by

$$Q_1 = \sum_{g=1}^{m} \sum_{i=1}^{n_1} e_{gi,1}^2 h_1(\theta_{1,i}).$$

(This approximation is made here for practical reasons; i.e., to reduce the computation time and the computer storage space needed.)

Moreover, it may be desirable that the criterion function also takes into account the effect of equating errors on the examinees whose ability parameters are defined on Scale 2. This can be achieved by first defining Q_2 by

$$Q_2 = \sum_{g=1}^{m} \sum_{i=1}^{n_2} e_{gi,2}^{\ 2} h_2(\theta_{2,i}),$$

where h_2 is the frequency distribution of θ_2, n_2 is the number of intervals, and

$$e_{gi,2} = P_{g,2}(\theta_{2,i}) - P_{g,1}(\theta_{21,i}),$$

where

$$\theta_{21,i} = (\theta_{2,i} - d)/k,$$

which defines a transformation of ability parameters on Scale 2 to Scale 1. Then the final form of the criterion function is given by

$$Q^* = Q_1 + Q_2$$

$$= \sum_{g=1}^{m} \left[\sum_{i=1}^{n_1} e_{gi,1}^{\ 2} h_1(\theta_{1,i}) + \sum_{i=1}^{n_2} e_{gi,2}^{\ 2} h_2(\theta_{2,i}) \right]. \qquad (8)$$

Similarly, the equating error for item g is

$$QI_g^* = \sum_{i=1}^{n_1} e_{gi,1}^{\ 2} h_1(\theta_{1,i}) + \sum_{i=1}^{n_2} e_{gi,2}^{\ 2} h_2(\theta_{2,i}). \qquad (9)$$

Equation 8 is recognized as the criterion function for a weighted least squares estimation of k and d with the weights being $h_1(\theta_{1,i})$ and $h_2(\theta_{2,i})$.

Solution

The weighted least squares solution for k and d is obtained by solving the simultaneous equations

$$\partial Q^*/\partial k = 0 \tag{10}$$

and

$$\partial Q^*/\partial d = 0 \tag{11}$$

for k and d. Equations 10 and 11 are to be solved numerically, since no closed form of analytical solution is available.

Among various techniques for solving such equations, we apply the Gauss-Newton method. For convenience of illustration, let us introduce some vector and matrix expressions:

$$\underline{x} = \begin{bmatrix} k \\ d \end{bmatrix}$$

$$\underline{f} = \left[P_{1,1}(\theta_{1,1}),\ P_{1,1}(\theta_{1,2}),\ldots,\ P_{m,1}(\theta_{1,n_1}),\right.$$
$$\left. P_{1,2}(\theta_{2,1}),\ P_{1,2}(\theta_{2,2}),\ldots,\ P_{m,2}(\theta_{2,n_2})\right]'$$

$$\hat{\underline{f}} = \left[P_{1,2}(\theta_{12,1}),\ P_{1,2}(\theta_{12,2}),\ldots,\ P_{m,2}(\theta_{12,n_1}),\right.$$
$$\left. P_{1,1}(\theta_{21,1}),\ P_{1,1}(\theta_{21,2}),\ldots,\ P_{m,1}(\theta_{21,n_2})\right]'$$

$\underline{W}$ = Diagonal matrix with elements

$$h_1(\theta_{1,1}),\ldots,\ h_1(\theta_{1,n_1}),\ h_2(\theta_{2,1}),\ldots,\ h_2(\theta_{2,n_2}).$$

Using these expressions, Equation 8 can be rewritten as

$$Q^* = (\underline{f} - \hat{\underline{f}})'\ \underline{W}(\underline{f} - \hat{\underline{f}}).$$

And Equations 10 and 11 are combined in the form

$$\underline{g}(\underline{x}) \equiv (\partial\hat{\underline{f}}/\partial\underline{x})' \underline{W}(\underline{f} - \hat{\underline{f}}) = \underline{0}.$$

Then the vector of estimates are updated successively by

$$\underline{x}^{(j+1)} = \underline{x}^{(j)} + t\underline{H}(\underline{x}^{(j)})^{-1}\underline{g}(\underline{x}^{(j)}), \tag{12}$$

where

$$\underline{H}(\underline{x}) = (\partial\hat{\underline{f}}/\partial\underline{x})' \underline{W}(\partial\hat{\underline{f}}/\partial\underline{x}),$$

and superscripts (j) and (j+1) refer to iteration numbers (see, for example, Dixon, 1972). A scalar t in Equation 12 is a step size factor which is used to keep the criterion function Q* from increasing. At the beginning of each iteration, t is set to one. And if Q* increases with this step size, then t is repeatedly halved until Q* decreases.

To implement this algorithm, one must specify the initial values for k and d. It is quite reasonable to use the values obtained by Equations 6 and 7 or by any other previously developed formulas. Or if one has some prior knowledge about the approximate values of k and d, he/she may use it.

APPLICATIONS

To illustrate the proposed method, equating was carried out for two sets of data.

Equating I

Applying the three-parameter logistic model with fixed non-zero c_g values, item parameters were estimated for a twenty-five-item

intelligence test (Cognitive Abilities Test, Verbal Battery, Test 1, Form 3, Level E), using a black seventh grade group (600 examinees) and a white seventh grade group (600 examinees), separately. Computer program LOGIST (Wood, Wingersky, & Lord, 1976) was used for the estimation of the item parameters. Then the scale for the black group was equated to the scale for the white group, using a computer program EQUATOR (Haebara, 1980). The initial values were computed by Equations 6 and 7.

Equating II

The test used in Equating II was a sixty-item achievement test (Tests of Achievement and Proficiency, Reading Comprehension Subtest, Form T, Level 15). In the same way as in Equating I, item parameters were estimated for a black ninth grade group (400 examinees) and a white ninth grade group (400 examinees), separately. Then the black group scale was equated to the white group scale as in Equating I.

Results and Discussion

The convergence processes in Equatings I and II are summarized in Tables 1 and 2, respectively. The convergence criterion adopted was the sum of the absolute values of changes in k and d being less than 10^{-4}. The last column of each table shows the values of an index of average error QM defined by

$$QM = (Q^*/2m)^{\frac{1}{2}}$$

where Q^* is defined by Equation 8. In the actual computation, only those examinees with ability parameter estimates between 3.1 standard deviations below and above the group mean were used in each group. This range

Table 1

Changes of Estimates and Criterion Function in Equating I (Intelligence Test)

Iteration	k	d	QM
Initial	.6135	-1.1785	.17525
1	.7659	-0.8831	.09228
2	.8601	-0.7510	.07175
3	.8919	-0.7294	.07066
4	.8971	-0.7279	.07064
5	.8977	-0.7279	.07064
6	.8978	-0.7279	.07064
7	.8978	-0.7279	.07064

Table 2

Changes of Estimates and Criterion Function in Equating II (Achievement Test)

Iteration	k	d	QM
Initial	1.2089	-.9770	.07793
1	1.1311	-.9432	.07703
2	1.1245	-.9402	.07702
3	1.1237	-.9399	.07702
4	1.1236	-.9398	.07702
5	1.1236	-.9398	.07702

was then divided into thirty-one intervals of size 0.2 to construct relative frequency distributions h_1 and h_2 in Equation 8.

The tables show rather quick convergence. In each equating only a few iterations were needed to reach the minimum value of QM to the fifth decimal place. A complete execution of EQUATOR took 7.44 seconds on an IBM 360/65 in Equating I and 12.86 seconds in Equating II.

The ratios of the final QM to the initial QM are .403 in Equating I and .988 in Equating II. In addition to the values of QM, the equating error for each item g was computed by

$$QMI_g = (QI_g^*/2)^{\frac{1}{2}},$$

where QI_g^* is defined by Equation 9. Using this index, it was found that the amount of error decreased, from the initial stage to the final stage of iterations, in twenty-two out of twenty-five items in Equating I, and in thirty-seven out of sixty items in Equating II.

The foregoing evaluation shows that the proposed method for equating worked more effectively in Equating I than in Equating II. This implies an advantage of the method: The big reductions of QM and QMI_g in Equating I are due to poor estimates of initial values (a previously proposed solution) for k and d caused by such extremely high b_g values as 169.0 and 115.6, and the present method seems to be more robust to the existence of such outliers and errors in the item parameter estimation than the previously developed method. This aspect of robustness is closely related to another advantage of the method. That is, those items with very high b_g values tend to have very low discriminating power as graphically represented by flat item characteristic curves, implying

little influence on the change of the criterion function and the result of equating. In other words, highly discriminating items tend to affect equating more strongly than less discriminating items.

A possible modification of the method to make it more robust to the existence of outliers may be to remove those items that have very high QMI_g values from the equating process. This may be achieved either at the computation of initial values or during the iteration process.

Incidentally, the value of the criterion function QM can be used as an index of goodness of fit of the logistic model to the items and the examinees involved. Similarly, QMI_g can be regarded as an index of the contribution of item g to the lack of fit of the model. Hence, removing those items with high QMI_g values not only improves equating but also helps to construct an item bank which the logistic model fits better.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Dixon, L. C. W. Nonlinear optimisation. London: English Universities Press, 1972.

Douglass, J. B. Applying latent trait theory to a classroom examination system: Model comparison and selection. Paper presented at the meeting of the American Educational Research Association, Boston, April 1980.

Haebara, T. A method for investigating item bias using Birnbaum's three-parameter logistic model (ITP Occasional Papers No. 25). University of Iowa, Iowa Testing Programs, December 1979.

Haebara, T. EQUATOR: A FORTRAN IV program for equating logistic ability scales by a weighted least squares method. Unpublished manuscript, 1980. (Available from the author.)

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eigner, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Shiba, S. Construction of a scale for acquisition of word meanings (Bulletin of Faculty of Education, Vol. 17, pp. 47-58). University of Tokyo, 1978. (In Japanese with abstract in English.)

Shiba, S., Noguchi, H., & Haebara, T. A stratified adaptive test of verbal ability. Japanese Journal of Educational Psychology, 1978, 26, 229-238. (In Japanese with abstract in English.)

Shiba, S., Noguchi, H., & Ohama, K. The efficiency of preliminary measurement by means of a short form of the stratified adaptive test in measuring verbal ability (Bulletin of Faculty of Education, Vol. 19, pp. 27-34). University of Tokyo. (In Japanese with abstract in English.)

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (ETS RM 76-6). Princeton, N. J.: Educational Testing Service, 1976.

IOWA TESTING PROGRAMS

OCCASIONAL PAPERS

1. Coffman, William E. Moratorium? What Kind? (ED 099 408)

2. Coffman, William E., & Mathews, Walter M. Narrative Reports and the Iowa Tests of Basic Skills.

3. Forsyth, Robert A., & Feldt, Leonard S. On the Validity of the I.T.B.D. as an Aid in Program Evaluation. (ED 142 558)

4. Gulliksen, Harold O. Applications of Psychological Scaling Methods.

5. Lee, Lai-Min Paul, & Coffman, William E. A Study of the "I Don't Know" Response in Multiple-Choice Tests. (ED 141 371)

6. Novick, Melvin R. A Course in Bayesian Statistics. (out of print) (In: American Statistician, May 1975, Vol. 29, No. 2, pp. 94-98)

7. Hood, Joyce, Leslie, Lauren, & Kendall, Janet Ross. Instructional Effectiveness in Primary Grade Reading: A Pilot Study.

8. Petersen, Nancy S., & Novick, Melvin R. An Evaluation of Some Models for Test Bias. (out of print) (In: Journal of Educational Measurement, Spring 1976, Vol. 13, No. 1, pp. 3-31)

9. Kendall, Janet Ross, & Hood, Joyce. Instructional Practices and Easy-to-Change Surrounding Conditions Variables in Effective Primary Reading Programs. (ED 123 557)

10. Petersen, Nancy S. An Expected Utility Model for "Optimal" Selection. (out of print) (In: Journal of Educational Statistics, Winter 1976, Vol. 1, No. 4, pp. 333-358)

11. Lindley, D. V. The Effect of Ethical Design Considerations on Statistical Analysis.

12. Lindley, D. V. A Class of Utility Functions. (In: Annals of Statistics, Vol. 4, No. 1, January 1976)

13. Lindley, D. V. Probability and Medical Diagnosis; and The Role of Utility in Decision-Making.

14. Lindley, D. V. Inference for a Bernoulli Process: A Bayesian View; and The Future of Statistics: A Bayesian 21st Century.

15. Shigemasu, Kazuo. Development and Validation of a Simplified M-Group Regression Model. (In: Journal of Educational Statistics, Summer 1976, Vol. 1, No. 2, pp. 157-180)

16. Cantor, Gordon N. Sex and Race Effects in the Community Behavior of Upper-Elementary-School-Aged Children. (ED 115 658)

17. Forsyth, Robert A. Describing What Johnny Can Do. (ED 181 063)

18. Sedere, M. U., & Feldt, Leonard S. The Sampling Distribution of the Kristof Reliability Coefficient, the Feldt Coefficient, and Guttman's λ_2. (In: Journal of Educational Measurement, Spring 1977, Vol. 14, No. 1, pp. 53-62)

19. Coffman, William E. Those Achievement Tests - What For?

20. Hunt, David E. Teachers are Psychologists, Too: On the Applications of Psychology to Education. (ED 129 651)

21. Johnson, Sylvia T. Self-Concept in a School Setting: Construct Validations by Factor Analysis. (ED 142 573)

22. Coffman, William E., & Shigemasu, Kazuo. Appraising School Effectiveness Using a Bayesian Method. (ED 159 220)

23. Coffman, William E. Classical Test Development Solutions. (ED 175 913)

24. Olsen, Scott A. An Investigation of Possible Test Floor Effects by Comparing Score Distributions for Students Tested with Two Different Levels of a Test Battery at Four Different Times over a Two-Year Interval. (ED 175 914)

25. Haebara, Tomokazu. A Method for Investigating Item Bias Using Birnbaum's Three-Parameter Logistic Model.

26. Coffman, William E., & Olsen, Scott A. The First Two Years of PLAN*--An Evaluation of Program Impact.

27. Haebara, Tomokazu. Equating Logistic Ability Scales by a Weighted Least Squares Method.